LARGE-SAMPLE SEQUENTIAL DECISION THEORY

EDWARD M. HOFSTETTER

TECHNICAL REPORT 359

DECEMBER 9, 1959

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the U. S. Army (Signal Corps), the U. S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039-sc-78108, Department of the Army Task 3-99-20-001 and Project 3-99-00-000.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 359 December 9, 1959

LARGE-SAMPLE SEQUENTIAL DECISION THEORY

Edward M. Hofstetter

Submitted to the Department of Electrical Engineering, August 24, 1959, in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

Statistical Decision Theory represents the latest attempt on the part of the statistician to formulate a general theory of experiment design. The theory allows for sequential experimentation, multivalued terminal decisions, and the use of several different types of experiment. An interesting and fairly broad class of decision problems can be conveniently described in the language of Communication Theory as follows: An experimenter is given a discrete, memoryless communication channel that is known to be one of a finite number of completely specified channels. His problem is to decide what sequence of input symbols to send over the channel and how to interpret the resulting output symbols in such a way as to yield an optimum procedure for guessing which channel is being used. An optimum procedure is understood to be one which, on the average, takes the shortest possible time commensurate with a given probability of guessing the wrong channel.

There exists a solution to this problem in the form of an iterative technique for determining the optimum experiment with any desired degree of accuracy. Since the number of iterations necessary to produce a good approximation is of the order of the expected duration of the experiment, this technique is useful only in small-sample situations. The present investigation, accordingly, is devoted to the development of a large-sample theory of experiment design. The main achievements of this theory are a limit theorem that describes the asymptotic behavior of the optimum experiment, and a specific experiment design that realizes this behavior. The theory sheds light on the behavior of sequential procedures, in general, and should prove useful in the solution of problems other than the specific one that is considered here.

TABLE OF CONTENTS

I. A CLASS OF DECISION PROBLEMS

1.1 INTRODUCTION

Statistical Decision Theory is a recently developed body of mathematics whose objective is to provide a rationale for the design of statistical experiments. The guiding principle of the theory is that the worth of a statistical experiment is determined solely by the ultimate use of the data derived from the experiment and cannot, therefore, be evaluated exclusively in terms of intrinsic properties of the experiment itself. Stated somewhat more mundanely, a given experiment may be of more value to experimenter A than to experimenter B because the experiment is "inaccurate" in just those situations in which A is not overly concerned with accuracy but B is, and vice versa. An analogous situation arises in Information Theory when the rate at which a stochastic source generates information depends on the combined characteristics of both the source and the information user and is not simply a property of the source itself. For a more detailed discussion of the philosophy of Statistical Decision Theory, the reader is referred to the excellent exposition given by Luce and Raiffa (1). The rest of Section I will be devoted to giving a precise formulation of the central problem of this report, in terms of which the concepts mentioned above will assume a concrete and, it is hoped, more understandable form.

1.2 THE CHANNEL N-CHOTOMY

An interesting and fairly broad class of decision problems can be stated as follows. A scientist wishes to determine something concerning the state of nature by performing a series of experiments. He has at his disposal a set of M different experiments each of which he can perform as many times as he sees fit. Corresponding to the j^{th} experiment is a set of possible outcomes X and a set of probability distributions over a set of possible outcomes $p_i^{(j)}(x)$, where i denotes the state of nature, and x is a member of X. To prevent purely technical problems from obscuring the issue, we shall assume that there are only N possible states of nature, and that the set X is finite. The scientist's problem is to determine which experiments to perform, in what order to perform them, and how to process the experimental data in such a manner as to arrive at an "accurate" estimate of i in the most "efficient" way. The words accurate and efficient will be given a precise meaning later.

The problem just described has received some attention in the literature of statistics (2, 3), but, thus far, no distinguishing name seems to have been coined for it. In this report we shall refer to the problem as the "Channel N-Chotomy" for reasons that will presently become apparent.

A discrete, memoryless communication channel, as defined in Information Theory (4), is a probabilistic device having a finite number of input and output symbols and a matrix of transition probabilities $p^{(j)}(x)$ that gives the probability that output

symbol x will occur, given that input symbol j was sent. The probability of a block of output symbols $x_1, x_2, \ldots, x_m$, given that the block of input symbols $j_1, j_2, \ldots, j_m$ was sent, is defined as $p^{(j_1)}(x_1)\, p^{(j_2)}(x_2) \ldots p^{(j_m)}(x_m)$. Such channels are often used as models for a noisy communication medium in which the transmission of a particular input symbol does not define a unique output symbol, but rather a probability distribution over the set of all possible output symbols.

In terms of the channel concept, the problem to be considered can be restated: We are given a discrete memoryless channel whose matrix of transition probabilities is unknown to us but which must be one of a finite number of known matrices $P_1, P_2, \ldots, P_N$. Our problem is to decide what sequence of input symbols to send, and how to interpret the resulting sequence of output symbols in such a manner as to yield an "accurate" and "efficient" procedure for determining the true value of the channel matrix. This restatement of the problem in the language of information theory may seem strange and unnatural to a reader who is unfamiliar with this discipline. However, doing so provides us with a model of the problem that is, in many ways, easier to talk about than the original.

Before attempting to solve our problem, we must, of course, give a precise mathematical description of the allowable procedures for determining the true value of the channel matrix, and we must assign precise meanings to the words "efficient" and "accurate" that are used in the statement of the problem. To this end, we define a test, or measuring, procedure for the channel N-chotomy to be a triple (g, s, d), where g denotes a go-ahead rule; s, a stopping rule; and d, a decision rule. A go-ahead rule is a collection of functions $\{g_o, g_1 \ldots\}$, where g_m maps the set of all sequences of channel output symbols of length m into the set of M channel input symbols. A stopping rule is a collection of functions $\{s_o, s_1 \ldots\}$, where s_m maps the set of all sequences of channel output sequences of length m into either 0 or 1. Finally, a decision rule is a collection of functions $\{d_o, d_1 \ldots\}$, where d_m maps the set of all sequences of channel output symbols of length m into the set of N possible channel matrices. These definitions still leave some confusion about the meaning of the symbols g_o, s_o, and d_o, whose function is to describe the beginning of the test. Accordingly, we define g_o to be any channel input symbol, s_o to be either 0 or 1, and d_o to be any one of the N possible channel matrices.

In terms of these definitions, the test procedure can be described inductively as follows: After the m^{th} channel output symbol has been received, we compute the value of $s_m(x_1, \ldots, x_m)$ and, according to whether this value is 0 or 1, we either decide to stop further testing and decide that channel matrix $d_m(x_1, \ldots, x_m)$ is present or to continue testing by sending channel input symbol $g_m(x_1, \ldots, x_m)$. In the latter case, the value of $s_{m+1}(x_1, \ldots, x_{m+1})$ is then computed, and the entire procedure repeated. The reader should note carefully that the tests that are being considered are sequential tests, in that the number of observations made is not fixed in advance but depends, in general,

on what the observed output symbols are. It will be appreciated that the concept of a test as defined above admits any reasonable scheme for extracting information from the unknown channel that we can possibly concoct.

Now that a precise description of the universe of tests to be considered has been given, we must decide on some basis for choosing one test in preference to another. For this purpose, we now introduce the concept of a loss function L(i, k), which is defined as a mapping of all pairs of possible channel matrices into the space of non-negative real numbers. L(i, k) is to be interpreted as a numerical index of our displeasure when we guess that channel k is present and then learn that channel i is actually present. A simple example of such a loss function is given by

$$L(i,k) = \begin{cases} 1, & i \neq k \\ 0, & i = k \end{cases} \tag{1}$$

which expresses the fact that all errors in guessing the true value of the channel matrix are equally undesirable. It should be made clear, now, that the selection of an appropriate loss function is not a statistical problem, but rather a reflection of the test designer's evaluation of how badly errors will affect the user of the data derived from the test.

Given any test $T = (g, s, d)$ and any loss function $L(i, k)$, we can define the expected loss of the test when channel i is present:

$$E_{iT}(L) = \sum_{k=0}^{\infty} \sum_{S_k} L[i, d_k(x_1, \ldots, x_k)]\ p_i(x_1, \ldots, x_k) \tag{2}$$

where

$$S_k = \{s_k(x_1, \ldots, x_k)=1, s_{k-1}(x_1, \ldots, x_{k-1})=0, \ldots, s_o=0\}$$

$$p_i(x_1, \ldots, x_k) = \text{probability of output sequence } x_1, \ldots, x_k, \text{ given that channel i is present}$$

and $E_{iT}(L)$ is a convenient measure of the average accuracy of test T when channel i is the true channel and will play an important part in the forthcoming analysis. Throughout this report the symbol {C} will be used to denote the set of all points satisfying condition C.

There is still another facet of test behavior that is of interest; this is the length of time it takes the test to come to the decision point. This length of time is not, in general, a fixed quantity, but rather a random variable $n(x_1, x_2, \ldots)$ which depends on the particular sequence of channel output symbols observed. In keeping with our definition of the expected loss of a test, we now define the expected length of a test T when channel i is present:

$$E_{iT}(n) = \sum_{k=0}^{\infty} \sum_{S_k} k\ p(x_1, \ldots, x_k) \tag{3}$$

We are now almost at the point where we shall be able to state precisely what it is that distinguishes a "good" test from a "bad" one. It should be obvious from definitions 2 and 3 that we are seeking tests that, in some sense, minimize expected length for a given expected loss. The difficulty involved in making such a definition is that pertinent expected lengths and losses all depend on which channel is actually present and, therefore, there may not, and usually there does not, exist a test that simultaneously minimizes <u>all</u> the $E_{iT}(n)$ for, say, a given set of $E_{iT}(L)$. This situation is similar to the one that arises in connection with the problem of uniformly most powerful tests in the theory of the Neyman-Pearson observer.

The difficulty completely disappears if we are willing to make the assumption that there exists an a priori probability distribution $\{\xi_i\}$ over the set of all possible channel matrices. In that case, we can define

$$E_T(L) = \sum_{i=1}^{N} \xi_i E_{iT}(L)$$

and

$$E_T(n) = \sum_{i=1}^{N} \xi_i E_{iT}(n)$$

and then say that an optimum test is one that minimizes $E_T(n)$ for a given $E_T(L)$. More precisely, an <u>optimum</u> test is one that minimizes $E_T(n)$, subject to the constraint that $E_T(L) \leqslant \bar{L}$, where $\bar{L}$ is a preassigned allowable loss level. The optimum tests so defined are referred to as "Bayes tests" for the channel N-chotomy.

Several serious objections to the use of a priori distributions in statistical problems have been raised, but the fact remains that the only really adequate criterion for choosing a "best" test is the one we have just given. A much stronger justification for discussing Bayes tests is contained in Wald's study of complete classes of tests, in which he shows that corresponding to any test that is not Bayes there is a Bayes test, which, in a certain sense, is better than the given test, regardless of which channel is actually present. We shall not make any attempt to pursue this highly controversial topic here, and merely refer the interested reader to either Luce and Raiffa (1) or Blackwell and Girschick (3) for a complete discussion of the issue.

The mathematical statement of the problem is now complete and can be described as the study of Bayes tests for the channel N-chotomy.

1.3 SMALL-SAMPLE THEORY

The main purpose of this section is to give a brief resume of what is known about the structure of Bayes tests for the channel N-chotomy. The main results in this direction are usually associated with the name of Wald, and the reader who is interested in a more detailed exposition of this subject than will be given here is referred to either Wald (2) or to Blackwell and Girschick (5).

The first result that we want to discuss gives a characterization of the Bayes tests which, in many ways, is more convenient to work with than the (g, s, d) characterization that we have used thus far. To state this result, we recall the well-known fact that any probability distribution over the set of N possible channel matrices can be represented as a point in an N-1 dimensional simplex S. It can be shown that every Bayes test corresponds to a division of S into (a) N convex stopping regions, each of which contains one, and only one, vertex of the simplex; and (b) a number of disjoint go-ahead regions, each labeled with some channel input symbol, whose union is equal to that part of the simplex external to the stopping regions. In terms of this division of S, the operation of the test is as follows: The point in S corresponding to the given a priori distribution $\{\xi_i\}$ is located. This point must lie either in a stopping region or in a go-ahead region. In the first case, the point lies in, say, the i^{th} stopping region, and the test is terminated with the decision that the i^{th} channel is present. In the second case, an observation is made by sending the channel input symbol corresponding to the pertinent go-ahead region, and then observing the resultant output symbol x_1. The a posteriori distribution given x_1, $\{\xi_{i1}\}$, is then computed from the formula

$$\xi_{i1} = \left[1 + \sum_{k \neq i} \frac{\xi_k}{\xi_i} \frac{p_k^{(j_1)}(x_1)}{p_i^{(j_1)}(x_1)} \right]^{-1}$$

where j_1 denotes the input symbol used, and the point in S corresponding to this distribution is located. The procedure just described is now repeated with the use of this new point instead of the point corresponding to the a priori distribution. It can be shown that not only does this procedure terminate with probability unity, but, in addition, the expected duration of the test $E_T(n)$ is finite. The finiteness of $E_T(n)$ is the only fact mentioned in this section that will explicitly be made use of hereafter.

The next result that we want to mention is closely related to the first, in that it describes an iterative procedure for determining the stopping and go-ahead regions with any desired degree of accuracy. The basic idea underlying this procedure is that a Bayes test can be regarded as the limit of a certain sequence of tests $\{T_m\}$, where T_m has the property that it terminates after no more than m observations have been made. The important point here is that the sequence $\{T_m\}$ can be constructed inductively, that is to say, T_o can be given explicitly, and T_m can be computed from T_{m-1}. Since T_m requires no more than m observations, it is obvious that the number of iterations necessary to produce a good approximation to T is of the order of $E_T(n)$. This property is one of the major drawbacks of the iterative procedure because it limits its application to small-sample situations. Another unpleasant feature of this technique is that the sequence of approximating tests $\{T_m\}$ depends on the a priori distribution, so that even if we had succeeded in determining the structure of a particular Bayes test, we should have to repeat the entire procedure if the a priori distribution were altered.

1.4 OUTLINE OF A LARGE-SAMPLE THEORY

In view of the difficulties involved in attempting to construct a Bayes test iteratively, it is natural to inquire whether there exist direct methods for designing tests which, although not themselves optimum, are, in some sense, close enough to optimum to be of use. The most reasonable way to attack this problem appears to be to attempt to exploit the familiar fact that large samples drawn from a statistical universe tend to behave in a strikingly regular manner. This regularity should evidence itself in our problem as a simplification of the structure of the Bayes tests when $E_T(n)$ is large. This idea was applied with great success by Wald (6) in his analysis of the channel dichotomy with one input symbol.

The rest of this report is devoted to the development of a large-sample theory of Bayes tests for the channel N-chotomy. The main achievements of this theory are a limit theorem that, in a well-defined sense, describes the asymptotic behavior of Bayes tests for large $E_T(n)$, and a specific test design that exhibits the same asymptotic behavior as the Bayes test. The theory sheds a great deal of light on the internal mechanics of sequential decision procedures, in general, and should provide some useful ideas pertinent to the solution of sequential problems other than the specific one considered here.

Before plunging into the mathematical details of the theory, we pause to make a few simplifying assumptions. First, we shall use only the particular loss function defined by Eq. 1 in the rest of this report. This loss function is used so frequently in statistics that its expected value is denoted by a special symbol, P_{eT}, and is referred to as the probability of error of test T. For reasons that will become clear in Section III, the asymptotic behavior of P_{eT} is the same as that for (almost) any other loss function, and hence restricting ourselves to a study of P_{eT} implies little loss of generality.

The next assumptions are aimed at ridding the theory of some pathological cases that otherwise might tend to obscure the main ideas involved. To this end, we assume that the channel transition probabilities have the property that $p_i^{(j)}(x) = 0$ if, and only if, $p_k^{(j)}(x) = 0$ for all k and, furthermore, we assume that there do not exist i, j and k, $i \neq k$, such that $p_i^{(j)}(x) = p_k^{(j)}(x)$ for all x. The first half of this assumption excludes from consideration all cases in which there is a nonzero probability of being able to definitely eliminate one, or more, of the N possible channels from consideration, after a finite number of observations. In other words, we are demanding that there be zero probability for an N-chotomy to become an (N-1)-chotomy, or lower, after a finite number of observations. Situations in which this is not the case obviously exhibit an anomalous large-sample behavior which, in extreme cases, may even result in the probability of error going to zero after a finite number of observations.

The second assumption states that all the channel symbols, when used singly, must be capable of separating the true channel from its environment of N possible channels.

In these words, this assumption looks completely different from the first one. However, the basic objective of both assumptions is to insure that the N-chotomy that is under consideration is really a situation involving N-fold ambiguity, and not one that can be decomposed into a collection of situations in which the degree of ambiguity is smaller than N.

The final justification for these assumptions is, of course, the fact that they enable us to prove some interesting theorems without unduly restricting their domain of validity. In this connection, it should be mentioned that our assumptions are of an almost purely technical character in that any N-chotomy that violates them can be transformed into one that does not, by means of a slight alteration of the pertinent channel matrices. This alteration can be effected, for example, by replacing zero elements of a channel matrix by, say, 10^{-100}, and equal elements by those that differ by, say, 10^{-100}. Thus our assumptions in no way affect the practical applications of the theory.

II. CHANNEL PROCESSES

2.1 DEFINITIONS

In order to study the asymptotic behavior of the Bayes tests for the channel N-chotomy, we must first obtain some information about the stochastic processes that arise in the problem. We define a channel process to be any stochastic process that can be generated by the application of a go-ahead rule to a discrete, memoryless channel.

The output symbols of a channel process are, of course, not independently distributed. However, the type of dependence exhibited by such processes is of a particularly simple form. The probability that an output symbol x_m occurs at time m is given by $p^{(j_m)}(x_m)$, the channel transition probability for some input symbol j_m. The particular input symbol to be used in this formula is determined by the past history of the process, $x_1, \ldots, x_{m-1}$, through the go-ahead rule. Thus, only one out of M possible distributions can obtain at any given time, and the go-ahead rule is the mechanism that determines which one. The probability of a block of m output symbols $x_1, \ldots, x_m$ is given by $p^{(j_1)}(x_1)\, p^{(j_2)}(x_2) \ldots p^{(j_m)}(x_m)$, where j_k depends on the output symbols $x_1, \ldots, x_{k-1}$.

To illustrate the breadth of this definition, we note that any finite-order, discrete, stationary Markov process is a channel process. For a k^{th}-order Markov process, the channel input symbols may be taken to be the set of all sequences of output symbols from the Markov process of length k, and the channel output symbols may be taken to be the set of output symbols from the Markov process. The channel transition probabilities are defined to be the Markov-process transition probabilities $p(x_m | x_{m-1}, \ldots, x_{m-k})$. The go-ahead rule consists in remembering the last k output symbols from the channel, and then choosing the corresponding channel input symbol. (A slight broadening of our definition of channel processes is necessary if we want the argument given above to encompass Markov processes whose initial state is chosen according to some a priori distribution.) The converse of this statement is not true; not every channel process is a finite-order Markov process.

2.2 A LIMIT THEOREM

Channel processes behave, in some respects, much as independent processes do. In particular, if f(x) is a real valued function of a single output symbol x of the process, it is possible to make rather strong statements about the behavior of $\frac{1}{m} \sum_{i=1}^{m} f(x_i)$ for large m. Before stating the theorem that we have in mind, we define a function $c_j(x)$ by means of the formula,

$$c_j(x) = \begin{cases} 1 & \text{if the input symbol that produced the output } x \text{ was symbol } j \\ 0 & \text{otherwise} \end{cases}$$

The notation $c_j(x)$ is somewhat deceiving because $c_j(x)$ is actually a function of the entire past history of the process, and is not a function of the single output symbol x. This slight abuse of notation will not, however, lead to any confusion in the sequel. In terms of this definition, we see that the expression

$$m_j = \sum_{i=1}^{m} c_j(x_i)$$

defines m_j as the number of times input symbol j was sent when output sequence $x_1, \ldots, x_m$ was received. If we are now given a set of M functions of a single output symbol $f_j(x)$, one for each of the possible input symbols, we can state the following theorem of Shannon (7).

THEOREM 1. Given any $\epsilon > 0$, there exists a $g > 0$ such that for all n,

$$P\left\{\left|\sum_{i=1}^{m}\sum_{j=1}^{M} c_j(x_i)\, f_j(x_i) - \sum_{j=1}^{M} m_j\, E_j(f_j)\right| > m\epsilon\right\} \leqslant 2e^{-gm}$$

where

$$E_j(f_j) = \sum_{X} p^{(j)}(x)\, f_j(x)$$

The proof of this theorem rests on the following lemma.

LEMMA 1. Let y denote a real, discrete, random variable whose range consists of the points $y_1, \ldots, y_N$ and whose expected value is zero. It follows that the moment-generating function of y, defined by

$$G(s) = \sum_{i=1}^{N} p(y_i)\, e^{y_i s} \qquad \text{for all real } s$$

satisfies the inequality $G(s) \leqslant G^*(s)$, where

$$G^*(s) = \tfrac{1}{2}e^{as} + \tfrac{1}{2}e^{-as}$$

and

$$a \geqslant \max_i |y_i|$$

Note carefully that $G^*(s)$ does not depend on $p(y_i)$.

PROOF OF LEMMA 1. For each i there exists a real number q_i, $0 \leqslant q_i \leqslant 1$, with the property that $y_i = q_i a + (1-q_i)(-a)$. It now follows from the convexity of the function e^{sy} that

$$e^{sy_i} = e^{s[g_i a-(1-g_i)a]} \leqslant g_i\, e^{as} + (1-g_i)\, e^{-as}$$

Therefore,

$$G(s) \leq \sum_{i=1}^{N} p(y_i)\left[g_i e^{as}+(1-g_i)e^{-as}\right]$$

The truth of lemma 1 now follows from the identities

$$\sum_{i=1}^{N} p(y_i)\, g_i = \frac{1}{a} \sum_{i=1}^{N} p(y_i)[y_i+(1-g_i)a] = \sum_{i=1}^{N} p(y_i)(1-g_i)$$

and

$$\sum_{i=1}^{N} p(y_i)\, g_i + \sum_{i=1}^{N} p(y_i)(1-g_i) = 1$$

PROOF OF THEOREM 1. We first note that the random variable defined by

$$g(x) = \sum_{i=1}^{M} c_j(x)[f_j(x)-E_j(f_j)]$$

has the property that for any m,

$$\sum_{X_m} g(x_m)\, p(x_m | x_{m-1}, \ldots, x_1) = 0 \qquad \text{for all } x_{m-1}, \ldots, x_1$$

This equation follows directly from the fact that the probability of x_m, given $x_{m-1}, \ldots, x_1$, is equal to $p^{(j_m)}(x_m)$, where the j_m depends on $x_{m-1}, \ldots, x_1$, through the go-ahead rule. We now compute the moment-generating function.

$$G(s) = \sum_{X_1, \ldots, X_m} p(x_1, \ldots, x_m)\, e^{sF_m}$$

where

$$F_m = \sum_{i=1}^{m} g(x_i)$$

It follows from lemma 1 that

$$G(s) = \sum_{X_1} p(x_1)\, e^{sg(x_1)} \cdots \sum_{X_m} p(x_m | x_{m-1}, \ldots, x_1)\, e^{sg(x_m)} \leq [G^*(s)]^m$$

where

$$G^*(s) = \tfrac{1}{2}e^{as} + \tfrac{1}{2}e^{-as}$$

and

$$a = \max_X |g(x)|$$

Next, we note that

$$G(s) = E\left[e^{sF_m}\right] \geq E\left[e^{sF_m} \mid U\right] P(U)$$

for any set of sequences U. In particular, we choose U to be the set defined by

$$U = \{F_m \geq m\epsilon\}$$

It now follows that

$$G(s) \geq e^{sm\epsilon}\, P(U), \qquad s \geq 0$$

or

$$P(U) \leq e^{g(s)-sm\epsilon}, \qquad s \geq 0$$

where $g(s) = \log G(s)$. Therefore

$$P(U) \leq e^{m[g^*(s)-s\epsilon]}, \qquad s \geq 0$$

where $g^*(s) = \log G^*(s)$.

A simple calculation now shows that $\frac{dg^*(0)}{ds} > 0$, from which it follows that an $s > 0$ can be found for which $g^*(s) - s\epsilon < 0$. Therefore, there exists a $g > 0$ for which

$$P\{F_m \geq m\epsilon\} \leq e^{-gm}$$

A repetition of the preceding argument, with the use of $-F_m$ instead of F_m, yields the result

$$P\{F_m \leq -m\epsilon\} \leq e^{-gm}$$

for some $g > 0$. Combining these two results completes the proof of theorem 1.

An important special case of theorem 1 – which we shall make use of in the sequel – is obtained by defining

$$f_j(x) = -\log p^{(j)}(x)$$

It follows directly from this definition that

$$\sum_{i=1}^{m} \sum_{j=1}^{M} c_j(x_i)\, f_j(x_i) = \log p^{-1}(x_1, \ldots, x_m)$$

and that

$$E_j(f_j) = -\sum_{X} p^{(j)}(x) \log p^{(j)}(x) = H^{(j)}$$

where $H^{(j)}$ denotes the entropy of the output process when only input symbol j is used.

Theorem 1 now reads

$$P\left\{\left|\frac{\log p^{-1}(x_1, \ldots, x_m)}{m} - \sum_{j=1}^{M} \frac{m_j}{m} H^{(j)}\right| > \epsilon\right\} \leq 2e^{-gm} \tag{4}$$

for some $g > 0$. The importance of this theorem lies in the fact that it gives us some precise information about how the go-ahead rule controls the behavior of the probability of blocks of output symbols when m is large.

2.3 CHANNEL PROCESSES WITH A STOPPING RULE

At this point, we must study how the addition of a stopping rule to a channel process affects the structure of the process. In particular, we shall be concerned with the behavior of random variables of the form

$$F_n = \sum_{i=1}^{n} \sum_{j=1}^{M} c_j(x_i)\, f_j(x_i)$$

where $c_j(x)$ and $f_j(x)$ are as defined in section 2.2, and $n = n(x_1, x_2, \ldots)$ denotes the random variable that gives the termination length of the test when the sequence of output symbols $x_1, x_2, \ldots$ is observed. Thus the only difference between the random variables to be considered in this section and those that were discussed in section 2.2 is that the number of terms in the summation is now a random variable. We can now state the following theorem.

THEOREM 2. If $E(n) < \infty$, then

$$E\left[\sum_{i=1}^{n} \sum_{j=1}^{M} c_j(x_i)\, f_j(x_i)\right] = \sum_{j=1}^{M} E(n_j)\, E_j(f_j)$$

where n_j denotes the number of times input symbol j was used before the process terminated. For notational convenience, we have dropped the subscript T from the pertinent expectation symbols.

This theorem was first proved by Wald (6) for the special case in which there is only one channel input symbol. The only new thing in our version of this theorem is the removal of this restriction.

PROOF OF THEOREM 2. We define

$$F_m = \sum_{i=1}^{m} c_j(x_i)[f_j(x_i) - E_j(f_j)]$$

for any integer, or any integral valued random variable, m. We choose an integer N and compute

$$E[F_N] = \sum_{i=1}^{N} E\{c_j(x_i)[f_j(x_i) - E_j(f_j)]\} = 0$$

in which the last equality follows directly from the definitions of the quantities involved.

The next step is to note that

$$E[F_N] = E[F_N | n \leq N]\, P\{n \leq N\} + E[F_N | n > N]\, p\{n > N\}$$

However,

$$\left| E[F_N | n > N] P\{n > N\} \right| \leq NA\, P\{n > N\}$$

where $A = \max_X |f_j(x) - E_j(f_j)| < \infty$. Since $E(n) < \infty$, it follows that

$$\lim_N N\, P\{n > N\} = 0$$

We can now conclude that

$$\lim_N E[F_N | n > N]\, P\{n > N\} = 0$$

and, furthermore, since $E[F_N] = 0$,

$$\lim_N E[F_N | n \leq N]\, p\{n \leq N\} = 0$$

We now introduce the random variable $F'_n = F_N - F_n$. Note that if $n \leq N$,

$$F'_n = \sum_{i=n+1}^{N} c_j(x)[f_j(x_i) - E_j(f_j)]$$

and, furthermore, the discreteness of the process and the fact that $E(n) < \infty$ imply that both $E[F_N]$ and $E[F'_n]$ exist. We now write

$$E[F_N | n \leq N] = E[F_n | n \leq N] + E[F'_n | n \leq N]$$

and note that the second term on the right vanishes. This follows from the fact that the random variable n depends only on the past of the output. In other words, if the sequence $x_1, x_2, \ldots$ is contained in the set $\{n \leq N\}$, then any other sequence $y_1, y_2, \ldots$ that agrees with $x_1, x_2, \ldots$ in the first $n = n(x_1, x_2, \ldots)$ places will also be contained in the set $\{n \leq N\}$. Therefore, for any fixed $m \leq N$, all possible values of $x_{m+1}, \ldots, x_N$ are used in forming the sum F'_m, and it follows that $E[F'_n | n \leq N] = 0$.

We have now shown that

$$\lim_N E[F_n | n \leq N]\, p[n \leq N] = 0$$

However, by definition, the left-hand side is equal to $E[F_n]$, and it follows that $E[F_n] = 0$. In other words,

$$E\left[\sum_{i=1}^{n} c_j(x_i)\, f_j(x_i)\right] = E(n_j)\, E_j(f_j)$$

If we now sum over j, we obtain theorem 2.

A particular version of theorem 2 that we shall need later can be obtained by setting $f_j(x) = -\log p^{(j)}(x)$. Theorem 2 then reads

$$E[\log p^{-1}(x_1, \ldots, x_n)] = \sum_{j=1}^{M} E(n_j)\, H^{(j)} \tag{5}$$

in which the same notation that was employed in section 2.2 is used. In Section III, Eq. 5 will be used to obtain a relationship between the probability of error and the expected length associated with any test for the channel N-chotomy.

Some additional insight into the meaning of theorem 2 can be obtained by pointing out its relationship to the theory of random walks. Consider, for this purpose, a one-dimensional walk defined as follows: We are given a discrete space X, a set of M probability distributions $p^{(j)}(x)$ over X, a set of M real valued functions $f^{(j)}(x)$ defined on X, and a mapping g(y) of the real line Y into the set of integers 1, 2, ..., M. The random walk starts from the origin by taking a step of length $f^{(j_1)}(x)$ with probability $p^{(j_1)}(x)$, where $j_1 = g(0)$. After m steps have been taken, an $m + 1^{th}$ step of length $f^{(j_m)}(x)$ is taken with probability $p^{(j_m)}(x)$, where $j_m = g(y_m)$, and y_m denotes the sum of the first m steps. In other words, this is a random walk in which both the size of the allowable steps and the probabilities with which these steps are taken depend on the location of the point from which the step is to be taken. It can be easily verified that the sequence $y_1, y_2, \ldots$ is a channel process. If we now agree to stop the walk the first time we progress A units, or more, to the right of the origin, or the first time we progress B units, or more, to the left of the origin, we shall have a channel process on which a stopping rule has been imposed. Theorem 2 can now be translated to yield the statement: If the average length of the random walk is finite, it is equal to the sum over j of the average length of a type-j step multiplied by the average number of times we took a type-j step. A type-j step is, of course, a step taken from a point y on the line for which g(y) = j. From this viewpoint, the truth of theorem 2 seems intuitively obvious.

III. ASYMPTOTIC BEHAVIOR

3.1 INTRODUCTION

Now that we have obtained some understanding of channel processes, we can go on to attack the main problem – the asymptotic behavior of Bayes tests. To accomplish this end, we shall first derive an asymptotic lower bound for the probability of error obtainable with any test having a finite expected length. After having established this bound, we shall then show that there exist tests which achieve it. The proof of this fact will be carried out by actually exhibiting a test that has the required asymptotic behavior. Since the proofs are somewhat lengthy, it is a good idea to state the final result in advance. If P_e and $E(n)$ denote the probability of error and the expected length of a given test (the subscript T has been dropped for notational convenience), then, as we shall show,

$$\lim \frac{\log P_e^{-1}}{E(n)} = I$$

as $E(n) \to \infty$, where I is some positive number that can be computed from the given channel parameters. An explicit formula for I will be included, together with the proof of this statement.

3.2 AN ASYMPTOTIC LOWER BOUND

Before formally stating and proving the theorems that we have in mind, we introduce some useful notation.

As part of the definition of a channel N-chotomy we are given N sets of channel transition probabilities, $p_i^{(j)}(x)$, $i = 1, \ldots, N$, $j = 1, \ldots, M$. In terms of these transition probabilities, we define the quantities

$$H_{ik}^{(j)} = -\sum_X p_i^{(j)}(x) \log p_k^{(j)}(x) \tag{6}$$

we shall refer to $H_{ik}^{(j)}$ as the i-k cross entropy for channel input symbol j. Next, we let $\{a_j\} = \{a_1, \ldots, a_M\}$ denote any probability distribution with M terms, and we define

$$I_i = \max_{\{a_j\}} \min_{k \neq i} \sum_{j=1}^{M} I_{ik}^{(j)} a_j \tag{7}$$

where

$$I_{ik}^{(j)} = H_{ik}^{(j)} - H_{ii}^{(j)}$$

In Section I, we stated that, in order to avoid certain obvious pathologies, we would only

consider channel N-chotomies for which $p_i^{(j)}(x) = 0$ if, and only if, $p_k^{(j)}(x) = 0$ for all k. If we make use of this fact, it follows that $H_{ik}^{(j)} < \infty$, and therefore $I_i < \infty$. Finally, it is an almost trivial exercise to show that $H_{ii}^{(j)} \leq H_{ik}^{(j)}$ and that the equality holds if, and only if, $p_i^{(j)}(x) = p_k^{(j)}(x)$ for all x. It now follows directly from the "separability" assumption of Section I that $I_{ik}^{(j)} > 0$ for all $k \neq i$ and all j, and, then, from this it follows that $I_i > 0$.

Now that these facts have been established, we can state the following theorem.

THEOREM 3. Given any $\epsilon > 0$ there exists an $A > 0$ such that any test for the channel N-chotomy that satisfies $A < E(n) < \infty$ also satisfies

$$\frac{\log P_e^{-1}}{E(n)} \leq I(1+\epsilon) \tag{8}$$

where

$$\frac{1}{I} = \sum_{i=1}^{N} \frac{\xi_i}{I_i}$$

PROOF OF THEOREM 3. We shall prove that relation 8 holds for the Bayes test associated with the loss level P_e. Since, for any given P_e, the Bayes test has the smallest possible E(n), it will follow that relation 8 holds for any test that has a finite expected length. Throughout this proof, subscripts on probabilities and expectations will refer to the fact that these quantities are to be computed from the probability distributions that obtain when the channel corresponding to the subscript is actually present. All quantities referring to a specific test, such as P_e, E(n), and so on, will be understood to refer to the optimum (Bayes) test.

To begin the proof, we define Q_i, for i = 1, . . ., N, to be the set of all (infinite) sequences of channel output sequences that lead to the decision that channel i is present. It then follows that

$$E_i\left[\frac{p_{kn}}{p_{in}}\middle| Q_i\right] = \frac{P_k(Q_i)}{P_i(Q_i)}$$

and (9)

$$E_i\left[\frac{p_{kn}}{p_{in}}\middle| \bar{Q}_i\right] = \frac{P_k(\bar{Q}_i)}{P_i(\bar{Q}_i)}$$

where $\bar{Q}_i$ denotes the complement of Q_i, $P(Q_i)$ denotes the probability of set Q_i when channel i is present, and p_{in} is a condensed notation for $p_i(x_1, \ldots, x_n)$. To avoid any possible misunderstandings that might arise from our notation, we remark that the symbol $E[p_{kn}/p_{in}|Q_i]$ denotes the sum of the quantity p_{kn}/p_{in} weighted by the probability $p_{in}/P_i(Q_i)$ over the set Q_i. The relationships Eqs. 9 are thus seen to be trivial consequences of the definition of conditional expectation.

Now, since a Bayes test has the property that E(n), and thus $E_i(n)$, are finite, we can apply theorem 2 in a form similar to that given by Eq. 5 to obtain

$$E_i\left[\log p_{kn}^{-1}\right] = \sum_{j=1}^{M} E_i(n_j)\, H_{ik}^{(j)} \tag{10}$$

From Eq. 10 it follows that

$$E_i\left[\log \frac{p_{kn}}{p_{in}}\right] = \sum_{j=1}^{M} E_i(n_j)\, I_{ik}^{(j)} \tag{11}$$

Next, we make use of Jensen's inequality, which states that for any random variable u, E (log u) $\leq$ log E(u), to obtain

$$E_i\left[\log \frac{p_{kn}}{p_{in}}\right] = E_i\left[\log \frac{p_{kn}}{p_{in}}\middle| Q_i\right] P_i(Q_i) + E_i\left[\log \frac{p_{kn}}{p_{in}}\middle| \bar{Q}_i\right] P_i(\bar{Q}_i)$$

$$\leq P_i(Q_i) \log E_i\left[\frac{p_{kn}}{p_{in}}\middle| Q_i\right] + P_i(\bar{Q}_i) \log E_i\left[\frac{p_{kn}}{p_{in}}\middle| \bar{Q}_i\right]$$

$$= P_i(Q_i) \log \frac{P_k(Q_i)}{P_i(Q_i)} + P_i(\bar{Q}_i) \log \frac{P_k(\bar{Q}_i)}{P_i(\bar{Q}_i)}$$

By means of some elementary inequalities, we can transform this expression to read

$$E_i\left[\log \frac{p_{kn}}{p_{in}}\right] \leq p_i(Q_i) \log P_k(Q_i) + \log 2$$

Next, we note the obvious relationships

$$P_{ek} \geq P_k(Q_i), \qquad k \neq i$$

and

$$P_{ei} = 1 - P_i(Q_i)$$

where, as usual, P_{ei} denotes the probability of error, given that channel i is present. Since the Bayes tests surely have the property that their expected length increases without bound as the probability of error goes to zero, it follows that given any $\epsilon > 0$, there exists an A with the property that any Bayes test satisfying $P(n) \geq A$ also satisfies $P_{ek} \leq \frac{\epsilon}{1+\epsilon}$, k = 1, ..., N. Therefore, we can write

$$E_i\left[\log \frac{p_{kn}}{p_{in}}\right] \leq \frac{1}{1+\epsilon} \log P_{ek} + \log 2$$

for $k \neq i$ and $E(n) \geq A$. This result can now be used to bound the left-hand side of Eq. 11 to yield

$$-\sum_{j=1}^{M} I_{ik}^{(j)} E_i(n_j) \leq \frac{1}{1+\epsilon} \log 2 P_{ek}$$

or

$$2 P_{ek} \geq \exp\left[-(1+\epsilon) \sum_{j=1}^{M} E_i(n_j) I_{ik}^{(j)}\right]$$

Since the last inequality holds for $k \neq i$, we can sum over i, with the result that

$$2 P_{ek} \geq \frac{1}{N-1} \sum_{i \neq k} \exp\left[-(1+\epsilon) \sum_{j=1}^{M} E_i(n_j) I_{ik}^{(j)}\right]$$

which can then be averaged with respect to the a priori distribution, which, in turn, yields the inequality

$$P_e \geq \frac{1}{2N} \sum_{k=1}^{N} \sum_{i \neq k} \xi_k \exp\left[-(1+\epsilon) \sum_{j=1}^{M} E_i(n_j) I_{ik}^{(j)}\right]$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \sum_{k \neq i} \xi_k \exp\left[-(1+\epsilon) E_i(n) \sum_{j=1}^{M} I_{ik}^{(j)} \frac{E_i(n_j)}{E_i(n)}\right]$$

which is valid for $E(n) > A$.

Now, we notice that if we delete all the terms in the $k \neq i$ summation except the one with the least negative exponent, we shall obtain an even smaller lower bound. Thus

$$P_e \geq \frac{\xi^*}{2N} \sum_{i=1}^{N} \exp\left[-(1+\epsilon) E_1(n) \min_{k \neq i} \sum_{j=1}^{M} I_{ik}^{(j)} \frac{E_i(n_j)}{E_i(n)}\right] \tag{12}$$

where $\xi^* = \min_i \xi_i$. (We tacitly assume throughout this report that the a priori distribution is not degenerate because, if it were, we would actually have an (N-1)-chotomy, or lower.)

Next, we minimize the i^{th} exponent in relation 12 with respect to the variables $E_i(n_j)/E_i(n)$, $j = 1, \ldots, M$, subject to the constraint $\sum_{j=1}^{M} E_i(n_j)/E_i(n) = 1$, and obtain the result

$$P_e \geq \frac{\xi^*}{2N} \sum_{i=1}^{N} \exp[-(1+\epsilon) I_i E_i(n)] \tag{13}$$

where I_i is the quantity defined previously.

The final step in the proof is to minimize the right-hand side of relation 13 with respect to the variables $E_i(n)$, $i = 1, \ldots, N$, subject to the constraint $\sum_{i=1}^{N} \xi_i E_i(n) = E(n)$. The details of this minimization problem are carried out in Appendix A, and the result given there is

$$P_e \geq \frac{\xi^*}{2N} e^{-(1+\epsilon)IE(n)}$$

where

$$\frac{1}{I} = \sum_{i=1}^{N} \frac{\xi_i}{I_i}$$

We have now shown that corresponding to any $\epsilon > 0$, there is an A with the property that $E(n) \geq A$ implies

$$\frac{\log P_e^{-1}}{E(n)} \leq (1+\epsilon) I + \frac{\log [2N/\xi^*]}{E(n)}$$

Since E(n) can be made arbitrarily large, the validity of theorem 3 follows.

3.3 THE EFFICIENCY OF A TEST

Theorem 3 is of fundamental importance in the design of tests for the channel N-chotomy because it gives a measure of the ultimate performance that can be expected from any test. Stated somewhat loosely, theorem 3 tells us that any test that achieves a probability of error P_e must require, on the average, at least $\log P_e^{-1}/I$ observations. Thus the quantity I provides an absolute standard of performance that can be used to judge the merit of any particular test for which E(n) is sufficiently large.

In the light of the preceding discussion, it seems reasonable to define the efficiency of a particular test to be the quantity

$$e_T = \frac{\log P_e^{-1}}{IE(n)}$$

where P_e and E(n) are to be computed with respect to the test T. Theorem 3 provides us with partial justification for the use of the word "efficiency" in connection with the quantity e_T because it states that no long test can have an efficiency essentially greater than unity.

3.4 AN ASYMPTOTIC UPPER BOUND

The question that naturally arises is whether or not there exist tests with efficiencies arbitrarily close to unity. An affirmative answer, providing full justification for the definition of efficiency, is given by the following theorem.

THEOREM 4. Given any $\epsilon > 0$, there exist tests with arbitrarily large expected lengths for which

$$\frac{\log P_e^{-1}}{E(n)} \geq I(1-\epsilon)$$

In particular, the Bayes test has this property, provided only that its expected

length is sufficiently large.

PROOF OF THEOREM 4. The method of proof will be to actually construct a test having the desired property. The test that we have in mind is defined by the following go-ahead, stopping, and decision rules.

(a) The Go-Ahead Rule

A solution of the maximin problem given by Eq. 7 consists of an integer k(i), and an M term probability distribution $\{a_{ij}\}$ having the property that

$$I_i = \sum_{j=1}^{M} I_{ik(i)}^{(j)} a_{ij}, \qquad i = 1, \ldots, N \tag{14}$$

The k(i) and a_{ij} defined by Eqs. 7 and 14 may not be unique, and this complicates the argument somewhat. If the solution is not unique, we pick for the a_{ij} to be used in the following argument any set that satisfies Eq. 14. Corresponding to this choice of the a_{ij}, there will be one or more integers k(i); we denote the set of all of these integers by S_i. This freedom of choice of the a_{ij} in the case of multiple solutions may actually be to our advantage, in some cases, because one set of a_{ij} may lead to a much simpler test than another, as will soon become apparent.

We are going to assume that the a_{ij} are all rational numbers, and therefore representable in the form $a_{ij} = r_{ij}/r$, where the r_{ij} and r are integers. In case the a_{ij} were not rational, they could be replaced by rationals without altering the right-hand side of Eq. 14 by more than any preassigned ϵ. The replacement of I_i by $I_i \pm \epsilon$ in no way alters the theorem we are trying to prove; therefore the assumption of rationality does not imply a loss of generality.

The go-ahead rule examines the a posteriori distribution $\{\xi_{im}\}$ at every r^{th} step, and then decides what the next r input symbols are to be in the following way. If ξ_{im} is the maximum component of $\{\xi_{im}\}$, then input symbol 1 is sent r_{i1} times, input symbol 2 is sent r_{i2} times, and, finally, input symbol M is sent r_{iM} times during the course of the next r observations. This works out exactly right because of the fact that $\sum_{j=1}^{M} r_{ij} = r$.

(b) The Stopping and Decision Rules

The stopping rule is much simpler to describe than the go-ahead rule. We select a number δ, $0 < \delta < 1$, and then agree to stop the test the first time any component of the a posteriori distribution $\{\xi_{im}\}$ rises above the threshold level $1 - \delta$. For reasons of analytical simplicity, we shall assume that the a posteriori distribution is only examined every r^{th} step so that the termination length of the test is always an integral multiple of r. The decision rule corresponding to this stopping rule consists of guessing that channel i is present if ξ_{im} is the component of $\{\xi_{im}\}$ that caused the test to stop.

We shall now proceed to show that as long as δ is chosen sufficiently small, the test just defined will have the desired behavior. The first fact to note in this connection is that P_e is obviously less than or equal to δ, as long as it can be shown that the

test terminates with probability one. That this is indeed the case follows from the fact that E(n) is finite; this will be demonstrated in the course of the proof.

The bulk of the following argument will be concerned with obtaining upper and lower bounds for probability ratios of the form p_{km}/p_{im}. The motivation for this is the fact that the a posteriori distribution $\{\xi_{im}\}$ can be written in the form

$$\xi_{im} = \left[1 + \sum_{k \neq i} \frac{\xi_k}{\xi_i} \frac{p_{km}}{p_{im}}\right]^{-1} \tag{15}$$

so that anything we know about the behavior of p_{km}/p_{im} can be used to tell us something about the behavior of the a posteriori distribution, and vice versa. In the discussion that follows, the integer i is some fixed integer between 1 and N.

The first fact that we need can be obtained as a special case of theorem 1. More precisely, it follows from theorem 1 that given any $\epsilon > 0$, there exists a $g > 0$ for which

$$P_i\left\{\left|\frac{\log p_{km}^{-1}}{m} - \sum_{j=1}^{M} \frac{m_j}{m} H_{ik}^{(j)}\right| > \epsilon\right\} \leq 2\, e^{-gm} \tag{16}$$

The technique used here is similar to that used in deriving Eq. 4.

If we start from Eq. 16, it is possible to show that

i. E(n) is finite.

ii. Given $\epsilon > 0$, there exists a $g > 0$ and an m_o with the property that $m > m_o$ implies

$$P_i\left\{\left|\frac{m_j}{m} - \frac{r_{ij}}{r}\right| > \epsilon\right\} \leq A\, e^{-gm}$$

where A is some constant. The g in this theorem can obviously be made equal to the g in Eq. 16, simply by using the smaller of the two in both places.

iii. Given any $\epsilon > 0$, there exists a $\delta_o > 0$ with the property that $\delta \leq \delta_o$ implies

$$\frac{E_i(n_j)}{E_i(n)} \geq (1-\epsilon)\frac{r_{ij}}{r}$$

The proofs of these statements are given in Appendix B.

We can now combine Eq. 16 and result (ii) to obtain the result that given any $\epsilon > 0$, there exists an m_o with the property that $m > m_o$ implies

$$P_i\left\{\left|\frac{\log p_{km}^{-1}}{m} - \sum_{j=1}^{M} H_{ik}^{(j)} \frac{r_{ij}}{r}\right| > \frac{\epsilon}{2}\right\} \leq B\, e^{-gm} \tag{17}$$

where B is some constant. If we let R_{im} denote the set of sequences of channel output symbols defined by

$$R_{im} = \bigcap_{k=1}^{N} \left\{ \left| \frac{\log p_{km}^{-1}}{m} - \sum_{j=1}^{M} H_{ik}^{(j)} \frac{r_{ij}}{r} \right| \leq \frac{\epsilon}{2} \right\} \tag{18}$$

it follows from relation 17 that $m \geq m_o$ implies

$$P_i(\bar{R}_{im}) \leq B\, e^{-gm} \tag{19}$$

where $\bar{R}_{im}$ denotes the complement of R_{im}.

We have succeeded, thus far, in showing that the set R_{im} contains an overwhelming majority of the channel output sequences if channel i is present, and if m is large. We now complete this phase of the analysis by showing that sequences in R_{im} behave in a very simple manner. Indeed, for any sequence in R_{im} and $m \geq m_o$, it follows that for any k,

$$\exp\left[-m \sum_{j=1}^{M} \left(I_{ik}^{(j)}+\epsilon\right) \frac{r_{ij}}{r}\right] \leq \frac{p_{km}}{p_{im}} \leq \exp\left[-m \sum_{j=1}^{M} \left(I_{ik}^{(j)}-\epsilon\right) \frac{r_{ij}}{r}\right] \tag{20}$$

from which it is easily seen that

$$\frac{\sum_{k\neq i} \frac{\xi_k}{\xi_i} \frac{p_{km}}{p_{im}}}{\frac{\xi_{k(i)}}{\xi_i} \frac{p_{k(i),m}}{p_{im}}} \leq 1 + \frac{\sum_{k\neq i,\, k(i)} \frac{\xi_k}{\xi_i} \exp\left[-m \sum_{j=1}^{M} \left(I_{ik}^{(j)}-\epsilon\right) \frac{r_{ij}}{r}\right]}{\frac{\xi_{k(i)}}{\xi_i} \exp\left[-m \sum_{j=1}^{M} \left(I_{ik(i)}^{(j)}+\epsilon\right) \frac{r_{ij}}{r}\right]}$$

$$= 1 + \sum_{k\neq i,\, k(i)} \frac{\xi_k}{\xi_{k(i)}} \exp\left[-m \sum_{j=1}^{M} \left(I_{ik}^{(j)}-2\epsilon\right) \frac{r_{ij}}{r} + mI_i\right] \tag{21}$$

for any k(i) in S_i. The last equality follows directly from the definitions of I_i and S_i. Those terms of the sum in Eq. 21 for which k is not in S_i all have negative exponents because for such values of k,

$$I_i < \sum_{j=1}^{M} I_{ik}^{(j)} \frac{r_{ij}}{r}$$

The remaining terms in the sum can be bounded above by an expression of the form $C \exp(2m\epsilon)$. It now follows that if $m > m_o$,

$$\sum_{k\neq i} \frac{\xi_k}{\xi_i} \frac{p_{km}}{p_{im}} \leq \frac{\xi_{k(i)}}{\xi_i} \frac{p_{k(i),m}}{p_{im}} [1+\epsilon+C\, e^{2m\epsilon}] \tag{22}$$

for all sequences in R_{im}. Inequality 22 can obviously be replaced by

$$\sum_{k \neq i} \frac{\xi_k}{\xi_i} \frac{p_{km}}{p_{im}} \leq D\, e^{2m\epsilon} \frac{p_{k(i),\,m}}{p_{im}} \tag{23}$$

where D is some positive constant. An analogous argument can be used to derive a lower bound similar to Eq. 23, except for a minus sign in front of the ϵ. In other words, the logarithm of the left-hand side of Eq. 23 divided by m behaves essentially like the logarithm of $p_{k(i),\,m}/p_{im}$ divided by m, provided only that the pertinent sequences of channel symbols are in the set R_{im}, and that m is large. This fact is the crux of our entire argument.

It is a direct consequence of Eq. 20 that the probability ratios p_{km}/p_{im} approach zero with increasing m for sequences in R_{im}. It follows that we can choose an m_o sufficiently large that $m \geq m_o$ implies that $\xi_{im} \geq \frac{3}{4}$ for all sequences in R_{im}. This is the last condition that we are going to impose on m_o, hence we can now choose the δ for our test sufficiently small to force the minimum length at which the test can terminate to be greater than m_o. This can always be done, since there are only a finite number of channel output symbols. As a result of this choice of δ, we can say that any (infinite) channel output sequence x that is contained in the set $R_{in(x)}$, where n(x) denotes the termination length of the test when x is the output sequence, has the property that it leads to the decision that channel i is the channel present.

We have now assembled enough information about the proposed test to enable us to begin the final phase of the proof. Our main tool for this purpose will be a special case of theorem 2, which reads

$$E_i\left[\log \frac{p_{k(i),\,n}}{p_{in}}\right] = -\sum_{j=1}^{M} E_i(n_j)\, I_{ik(i)}^{(j)} \tag{24}$$

The derivation of Eq. 24 follows the same lines as the derivation of Eq. 5.

We are now going to place bounds on both sides of Eq. 24, as follows: If x denotes an (infinite) channel output sequence that is contained in the set $R_{in(x)}$, then x leads to the decision that channel i is present, and it follows that $\xi_{in} \geq 1 - \delta$, and therefore that $\xi_{i,\,n-r} \leq 1 - \delta$. Equation 15 tells us that this implies that

$$\sum_{k \neq i} \frac{\xi_k}{\xi_i} \frac{p_{k,\,n-r}}{p_{i,\,n-r}} \geq \frac{\delta}{1-\delta}$$

which, after we have applied Eq. 23, yields

$$D\, e^{2(n-r)\epsilon} \frac{p_{k(i),\,n-r}}{p_{i,\,n-r}} \geq \frac{\delta}{1-\delta} \tag{25}$$

Inequality 25 can be rewritten in the form

$$\log \frac{p_{k(i),\,n-r}}{p_{i,\,n-r}} \geq \log \frac{\delta}{1-\delta} - \log D - 2(n-r)\,\epsilon \tag{26}$$

Since there are only a finite number of output symbols,

$$\ell_i = \min_{x,\,j} \log \frac{p^{(j)}_{k(i)}(x)}{p^{(j)}_{i}(x)}$$

exists, and we obtain

$$\log \frac{p_{k(i),\,n}}{p_{in}} \geq \log \frac{\delta}{1-\delta} - \log D + r\ell_i - 2(n-r)\,\epsilon \tag{27}$$

which can be rewritten in the form

$$\log \frac{p_{k(i),\,n}}{p_{in}} \geq \log \delta - 2n\epsilon + E \tag{28}$$

where E is some positive constant.

We can now bound the left-hand side of Eq. 24 by writing

$$\begin{aligned} E_i\left[\log \frac{p_{k(i),\,n}}{p_{in}}\right] &= \sum_{m=0}^{\infty} \left\{ E_i\left[\log \frac{p_{k(i),\,n}}{p_{in}} \,\middle|\, n = m,\ R_{im}\right] P_i\{n=m, R_{im}\} \right. \\ &\qquad \left. + E_i\left[\log \frac{p_{k(i),\,n}}{p_{in}} \,\middle|\, n = m,\ \bar{R}_{im}\right] P_i\{n=m, \bar{R}_{im}\} \right\} \\ &\geq \sum_{m=0}^{\infty} [\log \delta + E - 2m\epsilon]\, P_i\{n=m, R_{im}\} + B\ell_i \sum_{m=0}^{\infty} m e^{-gm} \\ &\geq \log \delta + F - 2\epsilon\, E_i(n) \end{aligned} \tag{29}$$

where F denotes some positive constant. In deriving relation 29 use was made of the easily verified fact that ℓ_i is negative.

The last step in the proof consists in applying result (iii) to the right-hand side of Eq. 24 to obtain the fact that, for small δ,

$$-\sum_{j=1}^{M} E_i(n_j)\, I^{(j)}_{ik(i)} \leq -(1-\epsilon)\, I_i\, E_i(n) \tag{30}$$

Relation 30 combined with relation 29 yields

$$\log \delta + F - 2\epsilon\, E_i(n) \leq -(1-\epsilon)\, I_i\, E_i(n) \tag{31}$$

The fact that $P_e \leq \delta$ can now be used to rewrite relation 31 in the form

$$\log P_e \leq -E_i(n)[(1-\epsilon)I_i - 2\epsilon] - F \tag{32}$$

which, we recall, is valid for all sufficiently small δ. Since ϵ is arbitrary, and since $E_i(n) \to \infty$ as $\delta \to 0$, it follows from relation 32 that we can find a δ_o with the property that $\delta \leq \delta_o$ implies

$$\frac{\log P_e^{-1}}{E_i(n)} \geq I_i(1-\epsilon) \tag{33}$$

Relation 33 holds for all i; therefore it follows that to every $\epsilon > 0$ there corresponds a $\delta_o > 0$ with the property that $\delta < \delta_o$ implies

$$\frac{\log P_e^{-1}}{E(n)} \geq I(1-\epsilon)$$

where

$$\frac{1}{I} = \sum_{i=1}^{N} \frac{\xi_i}{I_i}$$

COROLLARY 1. The Bayes tests have the property that

$$\frac{\log P_e^{-1}}{E(n)} \to I$$

as $E(n) \to \infty$.

COROLLARY 2. Given any loss function $L(i, k)$ with the property that $L(i, k) > 0$ if, and only if, $i \neq k$, the Bayes tests have the property that

$$\frac{-\log E(L)}{E(n)} \to I$$

as $E(n) \to \infty$.

PROOF OF COROLLARY 2. The proof follows immediately from the inequalities

$$\underline{L}\,\ell(i, k) \leq L(i, k) \leq \overline{L}\,\ell(i, k)$$

where

$$\overline{L} = \max_{i \neq k} L(i, k)$$

$$\underline{L} = \min_{i \neq k} L(i, k)$$

and $\ell(i, k)$ is defined by Eq. 1.

Analogs of the last two theorems hold for more general loss functions than we have considered (8). We omit the detailed proofs, since they entail only slight modifications of those already given.

IV. CONCLUDING REMARKS

4.1 COMMENTS ON THEOREM 4

In proving theorem 4 we constructed a test that was asymptotically optimum in the sense that its efficiency approached unity as its expected length increased. There is one other aspect of this test that deserves special attention; this is the fact that the test structure is independent of the a priori distribution. In fact, the entire test is characterized by the number δ which defines the stopping threshold, so that specifying the allowable probability of error of the test completely determines it. The Bayes test does not enjoy this independence of the a priori distribution, and thus we see that our asymptotically efficient test has at least the advantage of increased simplicity to compensate for the fact that it is not the best possible test. It is also gratifying to know that even though we have assumed the existence of an a priori distribution, it is possible to find tests whose behavior is near the optimum, in some sense, but whose structure is independent of this distribution. Viewed in this light, the objections that have been raised against the use of a priori distributions become much less serious than they would be if the structure of any "good" test depended very heavily on the particular a priori distribution chosen.

4.2 VARIABLE-DURATION CHANNEL SYMBOLS

The foregoing discussion has been concerned with the structure of Bayes tests, which, it will be recalled, were defined as those tests having the smallest expected length commensurate with a given probability of error. When defining the expected length of a test, we tacitly assumed that the length of time it takes to make a single observation of the channel is independent of which particular channel input symbol is used. In this section, we shall remove this restriction and consider channels for which a single observation with the use of channel input symbol j takes t_j units of time. (Another common usage is to refer to t_j as the "cost" of performing experiment number j.) The duration of a sequence of m observations is defined as

$$t = \sum_{j=1}^{M} t_j m_j$$

where, as usual, m_j denotes the number of times channel input j was used in the course of m observations. In accordance with this definition of test duration, we define a Bayes test to be any test that minimizes E(t) for a given value of P_e.

The large-sample behavior of this extended class of Bayes tests can be analyzed, theorem for theorem, by the methods developed in Sections II and III. In particular, it is an almost trivial exercise to verify that theorems 3 and 4 hold exactly as before if E(n) is replaced by E(t) and if the value of I is defined by

$$\frac{1}{I} = \sum_{i=1}^{N} \frac{\xi_i}{I_i}$$

where

$$I_i = \max_{\{a_{ij}\}} \min_{k \neq i} \sum_{j=1}^{M} \frac{I_{ik}^{(j)}}{t_j} a_{ij}$$

and a_{ij} is a probability distribution on j for each i.

4.3 SUGGESTIONS FOR FURTHER RESEARCH

The large-sample theory developed in this report has several important drawbacks and should be regarded only as a first step toward a complete understanding of the asymptotic behavior of sequential decision processes. The most serious drawback is the fact that the theory does not yield any information concerning how large the expected length of a test must be in order to insure that the behavior of the probability of error is close to the large-sample limit predicted by the theory. What is needed here are sharp upper and lower bounds (rather than only asymptotic upper and lower bounds) on the behavior of $\log P_e^{-1}$, or, better still, bounds on P_e itself. Also, it would be very worth while to attempt to determine the asymptotic behavior of P_e, in the hope that such a study would yield some simple tests whose large-sample P_e behavior was optimum. A successful investigation along these lines would constitute a fairly complete solution to the problem of large-sample test design.

An interesting question about which very little is known deals with the effect of truncating the length of a Bayes test by placing an upper bound on the number of observations to be made. Since no practical application of the tests described in this report can be made without some form of truncation, it is essential to determine to what extent truncation alters the probability of error and the expected length of the nontruncated test. A possible attack on this problem would be to attempt to determine the large-sample distribution on the termination length n. This information could then be used to determine, among other things, where to set the truncation level so as to obtain a truncated test whose behavior differs but little from that of the nontruncated test.

Another valuable line of research would be to extend the results of this report to channels with a continuum of output symbols. The interest in such channels is far from being purely academic; it stems from a desire to attack some of the many practical problems which, it seems, are most easily couched in terms of continuous distributions. Judging from the discrete theory, it appears that the hypotheses necessary to obtain continuous analogs of theorems 3 and 4 will take the form of restrictions on the manner in which the probability ratios $p_k^{(j)}(x) \Big/ p_i^{(j)}(x)$ approach infinity, when they do so. The case for which all of the probability ratios remain bounded seems to present no difficulties over and above the discrete case; but this type of restriction excludes from

consideration many of the really interesting distributions such as the Gaussian.

This report has been concerned exclusively with sequential tests, rather than with the older fixed-length tests that dominated the scene until the appearance of Wald's pioneering papers on sequential analysis. A fixed-length test is defined exactly as the tests described in Section I, except for the fact that the stopping rule is chosen in such a manner that the test always requires the same number of observations, regardless of what the results of these observations are. The (sequential) Bayes test that we have analyzed is obviously a better test than the fixed-length Bayes test, in the sense that the sequential test requires, on the average, fewer observations to achieve a given probability of error than the fixed-length test. However, since the sequential test is more difficult to implement than the fixed-length test, it is natural to ask how much better the sequential test is, so that we can judge whether the improved performance is worth the effort. No really definitive answers to this question have yet been obtained, except in the case of a one-input-symbol dichotomy (9). This special case indicates that the large-sample efficiency of the fixed-length test is considerably smaller than that of the sequential test, and there is every reason to suspect that this is true in general.

As a final remark, we should like to say a few words about some of the possible applications of the theory developed in this report. The original motivation for the present study of sequential decision theory came from a desire to investigate the problem of communication over noisy channels whose statistics are unknown at both the transmitter and receiver. A reasonable way of attacking this problem seems to be to divide the transmitting time into two parts: the first part to be used for measuring the channel statistics; and the second part to be used for transmitting information. By means of such an arrangement, the statistical knowledge gained as a result of the initial measurement can be used to combat the noise present when the information-bearing signal is sent. These considerations suggest that the study of ways and means for efficiently measuring channel statistics may perhaps provide valuable insight into some important, and still unsolved, communication problems.

Another more direct application arises in connection with the design of radar systems. The basic radar problem may be looked at as a channel-measurement problem, in which the channel statistics are determined by the ranges and velocities of the targets present and the channel input symbols are determined by the different waveforms which the radar can transmit. It is well known (10) that the shape of the transmitted waveform governs the radar's ability to resolve different targets. For example, a short pulse is excellent for discriminating between targets that are close in range, but very poor for targets that are close in velocity; and a long pulse has exactly the opposite effect. A great deal of effort has been expended on the problem of designing radar waveforms that are suitable for various purposes, but, apparently, very little consideration has been given to the possibility of making a group of waveforms available at the transmitter, and then letting the transmitter

decide, on the basis of the received data, which waveforms to employ, and in what order to employ them in order to best resolve the targets present. The theory developed in this report seems ideally suited for handling this problem.

Acknowledgment

I am very happy to have this opportunity to thank some of the generous people with whom I have come in contact during my stay at Massachusetts Institute of Technology. In particular, I should like to thank Professor Edward Arthurs for introducing me to decision theory, and for supervising and encouraging my research. I am also very grateful to Professor Peter Elias for his infinite patience, and for his unflagging willingness to answer the flood of questions that I put to him.

APPENDIX A

DETAILS OF THE PROOF OF THEOREM 3

We wish to minimize the expression

$$\sum_{i=1}^{N} e^{-a_i x_i}$$

subject to the constraint

$$\sum_{i=1}^{N} \xi_i x_i = 1$$

To this end, we use the method of Lagrange multipliers, and proceed to minimize the expression

$$\sum_{i=1}^{N} \left[e^{-a_i x_i} + \lambda \xi_i x_i \right]$$

The equations for the minimum are

$$-a_i e^{-a_i x_i} + \lambda \xi_i = 0, \qquad i = 1, \ldots, N$$

and their solution is

$$x_i = \frac{-1}{a_i} \log \left[\frac{\lambda \xi_i}{a_i} \right]$$

The value of λ is determined from the constraint equation by writing

$$\sum_{i=1}^{N} -\frac{\xi_i}{a_i} \log \left[\frac{\lambda \xi_i}{a_i} \right] = 1$$

from which it follows that

$$-\log \lambda \sum_{i=1} \frac{\xi_i}{a_i} = A + 1$$

where A is some constant. It now follows that

$$e^{-a_i x_i} = \frac{\xi_i}{a_i} B \exp \left[\frac{-1}{\sum_{i=1}^{N} \frac{\xi_i}{a_i}} \right]$$

and that the desired minimum is of the form

$$C \exp\left[\frac{-1}{\sum_{i=1}^{N} \frac{\xi_i}{a_i}}\right]$$

where B and C are constants.

It is now obvious that if the constraint equation is replaced by

$$\sum_{i=1}^{N} \xi_i x_i = \bar{x}$$

then the minimum is of the form

$$C \exp\left[\frac{-\bar{x}}{\sum_{i=1}^{N} \frac{\xi_i}{a_i}}\right]$$

APPENDIX B

DETAILS OF THE PROOF OF THEOREM 4

(a) Proof of Result (i)

Let S_{im} denote the set of channel output sequences defined by

$$S_{im} = \bigcap_{k=1}^{N} \left\{ \left| \frac{\log p_{km}^{-1}}{m} - \sum_{j=1}^{M} \frac{m_j}{m} H_{ik}^{(j)} \right| \leq \epsilon \right\}$$

It now follows from relation 16 that

$$P_i(\bar{S}_{im}) \leq 2N\, e^{-gm}$$

For any sequence in S_{im},

$$\frac{p_{km}}{p_{im}} \leq \exp\left[- \sum_{j=1}^{M} \left(I_{ik}^{(j)} - 2\epsilon \right) m_j \right]$$

and it is seen that the ratio p_{km}/p_{im}, for $i \neq k$, can be made as small as desired, uniformly over S_{im}, simply by choosing m sufficiently large. (Recall that the assumptions of Section I imply that $I_{ik}^{(j)} > 0$ if $i \neq k$.) In particular, the integer m_o can be chosen sufficiently large that sequences in S_{im} have the property that the a posteriori distribution $\{\xi_{im}\}$ satisfies $\xi_{im} > \frac{3}{4}$, if $m \geq m_o$. This value of m_o will be used in the proof of results (ii) and (iii). For the proof of result (i) we choose an m_o' large enough so that for $m > m_o'$, all sequences in S_{im} have the property that $\xi_{im} > 1 - \delta$, where $1 - \delta$ is the threshold level defined by the stopping rule for the test. As a consequence of this choice of m_o', it follows that if x denotes an output sequence that terminates the test at step m, and if $m > m_o' + r$, then x must be contained in the set $\bar{S}_{i,\, m-r}$ because all sequences in $S_{i,\, m-r}$ terminate the test at step $m - r$, or earlier. Therefore, we can write

$$P_i\{n=m\} \leq P_i(\bar{S}_{i,\, m-r}) \leq 2N\, e^{-gm}, \qquad m \geq m_o' + r$$

and use this fact to compute

$$E_i(n) = \sum_{m=0}^{\infty} m\, P_i\{n=m\} = \sum_{m=0}^{m_o'+r-1} m\, P_i\{n=m\} + \sum_{m=m_o'+r}^{\infty} m\, P_i\{n=m\}$$

$$\leq \sum_{m=0}^{m_o'+r-1} m\, P_i\{n=m\} + \sum_{m=m_o'+r}^{\infty} m\, e^{-g(m-r)} < \infty$$

from which it follows at once that E(n) is finite.

(b) Proof of Result (ii)

It is obvious that for $m \geq m_o' + r$,

$$P_i\{\xi_{im} < 1-\delta\} \leq P_i(\bar{S}_{im}) \leq 2N\, e^{-g(n-r)}$$

In other words, ξ_{im} converges in probability P_i to unity. It can be shown (11) that ξ_{im} converges almost everywhere (with respect to P_i) to some random variable, and it follows at once that this random variable must be the constant unity. From these facts, it follows that ξ_{im} eventually becomes the maximum component of $\{\xi_{im}\}$, and remains so for all succeeding m. It is now obvious from the definition of the go-ahead rule that m_j/m converges almost everywhere to r_{ij}/r (with respect to P_i). It now follows, of course, that ξ_{im} converges in probability P_i to r_{ij}/r. All that remains to be done is to estimate the rapidity of convergence. To do this, we define T_{im} as the set of all output sequences for which ξ_{im} is the maximum component of $\{\xi_{im}\}$. It follows that if $m \geq m_o$, $T_{im} \supset S_{im}$, and therefore that

$$P_i(T_{im}) \leq P_i(\bar{S}_{im}) \leq 2N\, e^{-gm}$$

Next, we define

$$V_{ik} = \bar{T}_{i,k-1} \cap T_{ik} \cap T_{i,k+1} \cap \cdots$$

which means that V_{ik} is the set of all sequences for which ξ_{im} becomes the maximum component of $\{\xi_{im}\}$ at the k^{th} step, and remains so for all succeeding steps. It is obvious that

$$P_i(V_{ik}) \leq 2N\, e^{-g(k-1)} \qquad \text{if } k \geq m_o + 1$$

and

$$V_{ik} \cap V_{ij} = \phi \qquad \text{if } k \neq j$$

Now, any sequence in V_{ik} satisfies

$$\frac{m-k}{m}\,\frac{r_{ij}}{r} \leq \frac{m_j}{m} \leq \frac{m-k}{m}\,\frac{r_{ij}}{r} + \frac{k}{m}$$

(m is assumed to be an integer multiple of r). This can be rewritten as

$$\left|\frac{m_j}{m} - \frac{r_{ij}}{r}\right| \leq \frac{k}{m}$$

Therefore, if a sequence is in V_{ik} and if $k \leq [m\epsilon]$ ($[m\epsilon]$ denotes the largest integer contained in $m\epsilon$), it follows that

$$\left| \frac{m_j}{m} - \frac{r_{ij}}{r} \right| \leq \epsilon$$

Result (ii) now follows from

$$P_i\left\{\left| \frac{m_j}{m} - \frac{r_{ij}}{r} \right| > \epsilon\right\} = \sum_{k=0}^{\infty} P_i\left\{\left| \frac{m_j}{m} - \frac{r_{ij}}{r} \right| > \epsilon, V_k\right\}$$

$$= \sum_{k=[m\epsilon]+1}^{\infty} P_i\left\{\left| \frac{m_j}{m} - \frac{r_{ij}}{r} \right| > \epsilon, V_k\right\}$$

$$\leq \sum_{k=[m\epsilon]+1}^{\infty} e^{-g(k-1)}$$

$$= \frac{e^{-g[m\epsilon]}}{1 - e^{-g}}$$

(c) Proof of Result (iii)

First, we define the set V_i to be $V_i = \bigcap_{k=1}^{\infty} V_{ik}$, and then we note that the fact that $\xi_{im} \to 1[P_i]$ implies that $P_i(V_i) = 1$. Therefore, result (iii) follows from

$$E_i(n_j) = E_i(n_j | V_i) = \sum_{k=1}^{\infty} E_i(n_j | V_{ik})\, P_i(V_{ik})$$

$$\geq \frac{r_{ij}}{r} \sum_{k=1}^{\infty} E_i(n-k | V_{ik})\, P_i(V_{ik})$$

$$= \frac{r_{ij}}{r} E_i(n) - \frac{r_{ij}}{r} \sum_{k=1}^{\infty} k\, P_i(V_{ik})$$

$$\geq \frac{r_{ij}}{r}\left[E_i(n) - \sum_{k=1}^{m_o-1} k - \sum_{k=m_o}^{\infty} k\, e^{-g(k-1)}\right]$$

$$= \frac{r_{ij}}{r}[E_i(n)-A]$$

where A is a constant independent of δ [and thus independent of $E_i(n)$].

References

1. R. D. Luce and H. Raiffa, Games and Decisions (John Wiley and Sons, Inc., New York, 1957), see especially pp. 309-316.
2. A. Wald, Statistical Decision Functions (John Wiley and Sons, Inc., New York, 1950).
3. D. Blackwell and M. A. Girschick, Theory of Games and Statistical Decisions (John Wiley and Sons, Inc., New York, 1954).
4. C. E. Shannon and W. Weaver, The Mathematical Theory of Communication (The University of Illinois Press, Urbana, Illinois, 1949).
5. D. Blackwell and M. A. Girschick, op. cit., p. 256.
6. A. Wald, Sequential Analysis (John Wiley and Sons, Inc., New York, 1947).
7. C. E. Shannon (private communication, March 1959).
8. D. Blackwell and M. A. Girschick, op. cit., p. 238.
9. E. Hofstetter, Asymptotic behavior of optimum fixed-length and sequential dichotomies, Quarterly Progress Report No. 53, Research Laboratory of Electronics, M.I.T., April 15, 1959, pp. 117-121.
10. P. M. Woodward, Probability and Information Theory (McGraw-Hill Book Company, Inc., New York, 1953).
11. J. L. Doob, Stochastic Processes (John Wiley and Sons, Inc., New York, 1953), p. 348.